The Pangeo Platform: Interactive Data Analytics for CESM

Kevin Paul
NCAR / CISL
Data Analytics

• Analysis
  • Been doing this forever!
  • Increases information density
  • e.g., Reduce 20 PB of CMIP data to 100 1M papers

• Analytics?
  • Some say it’s a synonym
  • Some say it implies the technology and methodology used in analysis
  • Comes from the business community ("Business analytics")
MapReduce

- Started by Google in their seminal 2004 paper
  Jeffrey Dean and Sanjay Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters.”

Query
What does “malneirophrenia” mean?

SEARCH

Dictionary (one computer)

A - Z

Answer
“The foul, bad state of mind after a nightmare.”

The Pangeo Platform
MapReduce

• Started by Google in their seminal 2004 paper
  Jeffrey Dean and Sanjay Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters.”

What does “malneirophrenia” mean?

“The foul, bad state of mind after a nightmare.”

Map! Reduce!

Distributed Dictionary (many computers)

Query: What does “malneirophrenia” mean?

Search: “The foul, bad state of mind after a nightmare.”
Beyond MapReduce...

• **2006:** Apache releases their own version of MapReduce (Hadoop)

• **2012:** Apache releases Spark to address limitations of MapReduce
  - More than just Map + Reduce
  - Provides a way of dealing with *distributed data objects*, without needing to know that the data is distributed at all!

```python
   d = DistributedDictionary()
   query = d.find("malneirophrenia")
   answer = query.compute()
```

- Very useful for distributed databases (tables, spreadsheets, etc.)!
- …Not so useful for Arrays!

Doesn’t do anything until here!
Beyond MapReduce…

- **2014:** Matt Rocklin (Anaconda Inc.) makes first commit to **Dask**
  - “Spark in Python”
  - Expands on Spark’s native database-like structures
    - Adds **Distributed Multidimensional Arrays**

- **2014:** Stephen Hoyer (Climate Corporation) starts **Xarray**
  - Provides an easy-to-use in-memory implementation of NetCDF-like data arrays
  - Builds on Dask Arrays
  - Add a lot of nice functionality to NetCDF-like data
    - Easy grouping data by month, season, etc.
    - Array access via coordinate values and array indices
    - Easily read/write from/to NetCDF
Beyond MapReduce…

- **2016**: Ryan Abernathey (LDEO, Columbia U) organizes PyAOS workshop
  - Brought about 12 people interested in Dask+Xarray together
  - “How do we make Dask+Xarray+Jupyter work better?”
  - Branded our community with the name Pangeo

- **2017**: Ryan Abernathey leads writing of the NSF EarthCube grant
  - Awarded in the summer of 2017
  - Start Date: September 1, 2017

• NSF-1740648: (LDEO, Columbia University)
  • PIs: Ryan Abernathey, Naomi Henderson, Richard Seager, Michael Tippett, Chiara Lepore
  • Sub-award to Matthew Rocklin (Anaconda, Inc.)

• NSF-1740633: (NCAR)
  • PIs: Kevin Paul, Joseph Hamman, Davide Del Vento, Ryan May (Unidata)

• Mission:
  • To improve the core functionality and integration of Xarray + Dask + Jupyter
  • To improve ease of deployment at HPC centers (e.g., NCAR) and in commercial cloud
So, what is “Pangeo”?

• It’s a **Community**:
  • Documentation, Tutorials, etc.: [http://pangeo-data.org](http://pangeo-data.org)
  • Gitter Developer Chats: [https://gitter.im/pangeo-data/Lobby](https://gitter.im/pangeo-data/Lobby)

• It’s an NSF EarthCube **Project**:
  • NSF-1740648, NSF-1740633

• It’s a software **Stack** for data analysis:
  • Dask, Xarray, Jupyter Notebook/Lab/Hub

• It’s a **Platform** for data analysis at many locations
  • With deployments on GCP, AWS, Cheyenne and many others

NCAR | The Pangeo Platform
PANGEO

“A community platform for Big Data geoscience”

jupyter
Interactive portable environments

xarray
Easy-to-use multidimensional arrays for targeted to the geosciences

DASK
Parallel “Big Data Analytics”
DEMO

http://pangeo.pydata.org